OPEN ACCESS Freely available online

PLOS ONE

# Information Filtering in Sparse Online Systems: Recommendation via Semi-Local Diffusion

**Wei Zeng[1,2], An Zeng[2]*, Ming-Sheng Shang[1,3]*, Yi-Cheng Zhang[1,2,3]**

**1** Web Sciences Center, University of Electronic Science and Technology of China, Chengdu, People's Republic of China, **2** Department of Physics, University of Fribourg, Fribourg, Switzerland, **3** Institute of Information Economy, Hangzhou Normal University, Hangzhou, People's Republic of China

## Abstract

With the rapid growth of the Internet and overwhelming amount of information and choices that people are confronted with, recommender systems have been developed to effectively support users' decision-making process in the online systems. However, many recommendation algorithms suffer from the data sparsity problem, i.e. the user-object bipartite networks are so sparse that algorithms cannot accurately recommend objects for users. This data sparsity problem makes many well-known recommendation algorithms perform poorly. To solve the problem, we propose a recommendation algorithm based on the semi-local diffusion process on the user-object bipartite network. The simulation results on two sparse datasets, Amazon and Bookcross, show that our method significantly outperforms the state-of-the-art methods especially for those small-degree users. Two personalized semi-local diffusion methods are proposed which further improve the recommendation accuracy. Finally, our work indicates that sparse online systems are essentially different from the dense online systems, so it is necessary to reexamine former algorithms and conclusions based on dense data in sparse systems.

**Citation:** Zeng W, Zeng A, Shang M-S, Zhang Y-C (2013) Information Filtering in Sparse Online Systems: Recommendation via Semi-Local Diffusion. PLoS ONE 8(11): e79354. doi:10.1371/journal.pone.0079354

**Editor:** Angel Sánchez, Universidad Carlos III de Madrid, Spain

**Received** July 9, 2013; **Accepted** September 28, 2013; **Published** November 18, 2013



**Funding:** This work is supported by the opening foundation of Institute of Information Economy in Hangzhou Normal University (Grant No. PD12001003002002) and the Sichuan Provincial Science and Technology Department (Grant No. 2012FZ0120). W.Z. acknowledges the support from Sino-Swiss Science and Technology Cooperation Program (EG57-092011). A.Z. acknowledges the support from China Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: an.zeng@unifr.ch (AZ); shang.mingsheng@gmail.com (M-SS)

## Introduction

Owing to the rapid development of the Internet, people are confronted with abundant online contents, which makes it very time-consuming to select the needed information. This is often refereed as the information overload problem. In order to solve it, search engines and recommender systems are widely investigated and applied to real systems. The search engine returns the relevant contents based on the keywords given by users. Compared to the search engine, the recommender system provides personalized services for users by predicting the potential interests based on their historical choices.

Up to now, many recommendation algorithms have been proposed such as collaborative filtering (CF) [1–3], content-based analysis [4] and spectral analysis [5]. The matrix factorization algorithms have also been widely investigated by combining high scalability with predictive accuracy [6–7]. Recently, some physical processes, including mass diffusion [8,9], heat conduction [10] and electric circuit analysis [11], have been applied to design recommendation algorithms. The hybridization of the mass diffusion and heat conduction algorithm is shown to effectively solve the diversity-accuracy dilemma in recommendation [12]. Based on these algorithms, many methods have been proposed to further enhance the recommendation diversity and solve the object cold-start problems. For example, the preferential diffusion [13], the biased heat conduction [14], network manipulation [15] and the item-oriented method [16] are shown to be able to largely improve the recommendation accuracy for small-degree objects. More recently, the long-term influence of the hybrid approach on network evolution has been studied [17].

One of the biggest challenges in recommender systems is the data sparsity problem. That is, the user activity data is too sparse for the recommender system to provide satisfactory recommendations. To solve such sparsity problem, the users' social network is incorporated in the object recommendation. For instance, a random walk model based on both the trust network and user-object bipartite network was designed [18]. Based on the matrix factorization method, both the user trust network and friendship network can be fused in the object recommendation by regularization [19,20]. Yang [21] proposed a factor-based random walk model to recommend both online services and friends to users. In addition, the users' membership data (i.e. the social groups that online users joined) is considered and the results indicate that this social information is more valuable than friendship when used to enhance the recommendation accuracy of object [22].

However, the users' social network is usually much sparser than the user-object network in most systems. More importantly, those users who have collected or purchased few objects might also be inactive in building their social relationships. Therefore, the compensation effect of social networks on the user-object bipartite networks is limited. In this paper, we propose an approach based on the semi-local diffusion process on the user-object bipartite network to solve the data sparsity problem. Our simulation results on two real datasets, Amazon and Bookcross, indicate that our method significantly outperforms the state-of-the-art methods

especially for these small-degree users. Moreover, two personalized semi-local diffusion methods are proposed which further improve the accuracy.

## Data Sparsity Problem

An online commercial system can be usually represented by a bipartite network $G(U,O,E)$, where $U=\{u_1,u_2,...,u_N\}$, $O=\{o_1,o_2,...,o_M\}$ and $E=\{e_1,e_2,...,e_L\}$ are the sets of users, objects and links, respectively. Denote by an adjacency matrix $A$, where the element $a_{i\alpha}=1$ if user $i$ has collected item $\alpha$, and 0 otherwise (throughout this paper we use Greek and Latin letters, respectively, for item- and user-related indices) [2,23].

The hybrid method in ref. [12] takes into account both the mass diffusion [8] and the heat conduction [10] process. This method is shown to be able to provide not only accurate but also diverse recommendations for users when applied to dense datasets. Here, we argue that this hybrid method fails in sparse datasets. As an example, we test this hybrid method on two sparse datasets: Amazon (www.amazon.com) and Bookcross (www.bookcrossing.com). *Amazon.com* is a multinational e-commerce company and the world's largest online retailer. The original data was collected from 28 July 2005 to 27 September 2005 [24]. During this period, there are 1,714,512 reviewers in total. The data contains 100,000 highest ranked reviewers and all reviews written by them. Some of the reviewers in the list didn't give reviews during this period of time, so that in practice only 99,622 reviewers contributed. They wrote total 2,036,091 reviews on 645,056 products. Here, we select a random subset from the data. *Bookcrossing.com* is a book sharing web site where book lovers can exchange their books and experiences with each other. The original data has 278, 858 users and 1, 157, 112 ratings, referring to 271, 379 distinct ISBNs (objects) [25]. Invalid ISBNs were excluded from the dataset. The complete BookCrossing dataset is available online (http://www.informatik.uni-freiburg.de/~cziegler). The data in this paper is a random sample from the original data. Some basic statistics of these two datasets are presented in the Table 1. Each data is randomly divided into two parts: the training set ($E^T$) and the probe set ($E^P$). The training set contains 80% of the original links and the recommendation algorithm runs on it [26]. The rest of the links forms the probe set, which will be used to examine the recommendation performance.

When recommending objects for user $i$, the hybrid method works by assigning each object collected by user $i$ one unit of resource. The initial resources are denoted by the vector $\overrightarrow{f}$ where $f_\alpha$ is the resource possessed by object $\alpha$. Then they will be redistributed via the transformation $\overrightarrow{f'}=W\overrightarrow{f}$, where

$$W_{\alpha\beta}=\frac{1}{k_\alpha^{1-\lambda}k_\beta^{\lambda}}\sum_{j=1}^{N}\frac{a_{j\alpha}a_{j\beta}}{k_j} \quad (1)$$

**Table 1.** The statistics of Amazon and Bookcross datasets.

| Dataset | #user | #objects | #links | sparsity |
|---|---|---|---|---|
| Amazon | 50000 | 54,152 | 283,382 | $1035\times10^{-4}$ |
| Bookcross | 21122 | 203,373 | 504,643 | $1.17\times10^{-4}$ |

The sparsity is obtained by $\frac{\#links}{N\times M}$, where $N$ and $M$ are the number of users and items, respcetively.

doi:10.1371/journal.pone.0079354.t001

is the redistribution matrix, with $k_\alpha=\sum_{l=1}^{N}a_{l\alpha}$ and $k_j=\sum_{\gamma=1}^{M}a_{j\gamma}$ denoting the degree of object $\alpha$ and user $j$, respectively. $N$ and $M$ are the number of users and objects, respectively. $\lambda$ is a tunable parameter which adjusts the relative weight between the Mass Diffusion algorithm (short for MD, $\lambda=1$) and Heat Conduction algorithm (short for HC, $\lambda=0$). The illustration of MD and HC algorithms can be seen in Fig. 1(a) and (b), respectively. The resulting recommendation list of uncollected items is sorted according to $\overrightarrow{f'}$ in descending order.

In order to measure the recommendation accuracy, we make use of the ranking score ($RS$). Specifically, $RS$ measures whether the ordering of the items in the recommendation list matches the users' real preference. For a target user $i$, all her/his uncollected items will be ranked according to their predictive scores in the descending way by the recommender system. Given $\alpha$ is an object selected by user i in the probe set, $RS_{i\alpha}$ is the rank of $\alpha$ in $i$'s recommendation list divided by the total number of uncollected items by user $i$. The smaller the $RS_{i\alpha}$, the better the recommendation, the items in the probe set being ranked higher. The mean value of the $RS_{i\alpha}$ over all the user-item relations in the probe set can be used to evaluate the recommendation accuracy as

$$\langle RS\rangle=\frac{1}{|E^P|}\sum_{i\alpha\in E^P}RS_{i\alpha} \quad (2)$$

The smaller the value of $\langle RS\rangle$, the higher the recommendation accuracy.

In ref. [12], $\langle RS\rangle$ can achieve an optimal value when adjusting the parameter $\lambda$ of the hybrid recommendation method. However, when applied to the sparse data mentioned above, $\langle RS\rangle$ changes monotonously with $\lambda$, as presented in Fig. 2. In other words, the recommendation accuracy cannot be improved by taking into account the heat conduction process in the mass diffusion method.

To understand the reason, we introduce a concept called *coverage*, $c$. As shown in Fig. 1, the diffusion-based algorithms are based on 3 steps. Given the diffusion starting from user $i$, we denote the objects whose received resources are larger than 0 after

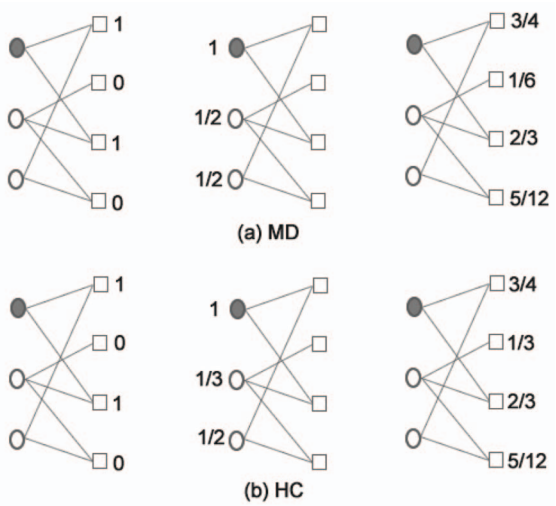


**Figure 1. The Mass Diffusion (a) and Heat Conduction (b) algorithms at work on the bipartite user-object network.** Users are shown as circles; objects are squares. The target user is indicated by the shaded circle.

doi:10.1371/journal.pone.0079354.g001

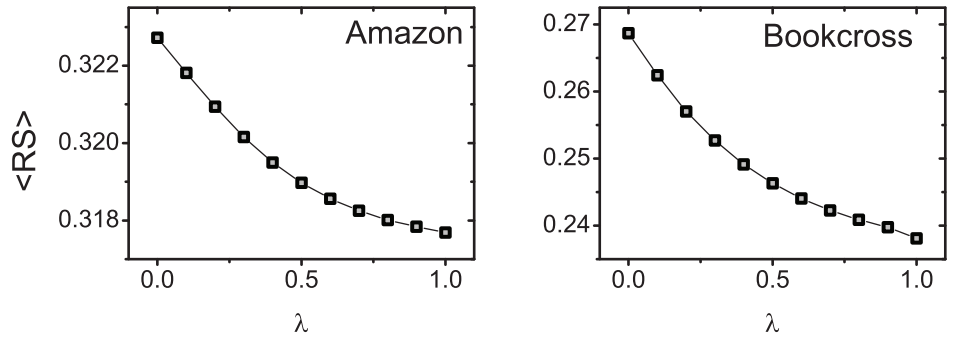


**Figure 2. The ranking score of the hybrid method on Amazon and Bookcross.** λ is used to tune the contribution of the heat conduction and the mass diffusion process. When λ = 1, the hybrid method gives the pure mass diffusion method and λ = 0 it degenerates to pure heat conduction method (more details about the hybrid method can be found in [12]). Each data point is obtained by averaging over ten runs, each of which has an independently random division of training set and probe set.
doi:10.1371/journal.pone.0079354.g002

3 diffusion steps as covered objects. Then the coverage $c_i$ is defined as the number of covered objects divided by the number of unselected objects by user $i$. Actually, this definition has been used before in [27]. The larger $c_i$ is, the more objects will receive resources in the Hybrid method. The average coverage $\overline{c}$ over all users are 0.0301 for Amazon and 0.1413 for Bookcross, respectively. In other words, most objects will receive 0 resource if we choose the Hybrid algorithm. Note that the hybridization [12] only changes the amount of resource of the covered objects. The resource of the uncovered objects will stay 0 under all hybrid parameters. Since the coverage dominates the recommendation accuracy in sparse data, the hybrid method cannot improve the recommendation accuracy as shown in Fig. 2. Moreover, we show the relationship between the user degree and the coverage $c$ in the top subfigures of Fig. 3. The coverage nonlinearly increases with user degree, which leads to an even more serious user cold-start problem. In next section, we will propose a semi-local diffusion method to increase the diffusion coverage and break the tie among these items with 0 resource.

## Algorithm and Metrics

Our semi-local diffusion method will be directly built on the mass diffusion method [8]. The MD method is simply the case when $\lambda = 1$ in the hybrid method. Given a target user $i$, the first step of MD is to allocate one unit resource to each of $i$'s collected items. Due to the bipartite structure, it takes two diffusion steps for the resource to get back to the item side. For convenience, we denote every 2 steps after the 1-step as one macro-step ($MS$ for short) of diffusion. The original 3-step diffusion is combined by the first ordinary step (the initial resources allocating process) and 1 macro-step diffusion. As discussed above, the original 3-step diffusion method suffers from the data sparsity problem since most objects' resources are 0. To solve this problem, we let the resources diffuse on the bipartite network more than one macro-step. The initial resources are denoted by the vector $\overrightarrow{f}$. After one macro-step, items' resource can be expressed as $\overrightarrow{f}^{(1)} = W\overrightarrow{f}$, where $W$ is the resource redistribution matrix for mass diffusion algorithm (with $\lambda = 1$ in equation 1). Likewise, we can calculate items' resource after $n$ macro-steps of diffusion as $\overrightarrow{f}^{(n)} = W\overrightarrow{f}^{(n-1)} = W^n\overrightarrow{f}$. To recommend objects to user $i$, one can sort the $\overrightarrow{f}^{(n)}$ in descending order and those objects with most resources will be recommended. Since the algorithm above uses less than global information but a bit more than pure local information, we call this method as Semi-Local Diffusion (SLD) recommendation method.

In previous section, we used the ranking score to measure the recommendation accuracy. Since real users usually consider only the top part of the recommendation list, a more practical measure should take into account the number of a user's hidden links contained in the top-L places. Therefore, we use another recommendation accuracy measure called "Recall". As discussed above, the real data is first divided into two parts: training set and probe set. For each user $i$, he/she may have certain number of links (corresponding to objects) in the probe set, we denote it as $E_i$. After the recommendation list (with length $L$) is generated for user $i$, we will calculate $d_i(L)$ as the number of his/her probe set objects which appear in the recommendation list. The Recall of this user is defined as

$$Re_i(L) = d_i(L)/E_i. \tag{3}$$

The Recall of the whole system is defined as

$$Re(L) = \frac{1}{N}\sum_{i=1}^{N} Re_i(L). \tag{4}$$

A higher Recall value indicates a higher accuracy of recommendation.

## Results

If we let the objects' resources diffuse on the bipartite network for multiple macro-steps, more objects will be covered. We plot the relations between the average coverage $\overline{c}$ and the macro-step in the bottom two subfigures of Fig. 3. As one can see, the average coverage $\overline{c}$ increased quickly with the macro-step. Therefore, more objects in the probe set may receive resource in the diffusion and have higher rank accordingly. The relation between the overall $\langle RS \rangle$ and the number of macro-steps is presented in Fig. 4. If *macro-step = 1*, the method degenerates to the standard Mass diffusion method. From the figure, one can see that $\langle RS \rangle$ is improved significantly by the SLD method and the optimal macro-step is 5 in both datasets. If the macro-step is more than 5, the ranking score gets worse but still much better than that of the original MD method. We actually test the other diffusion methods based on one macro-step [13,14,16], and the results show that

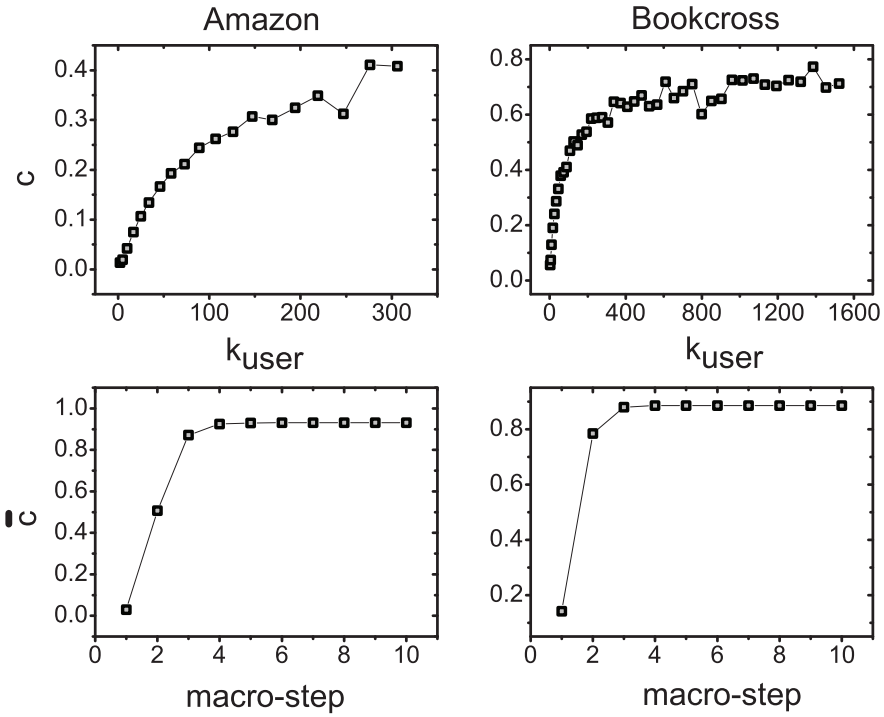


**Figure 3. The coverage $c$ and avgerage coverage $\bar{c}$ in Amazon and Bookcross.** The top two subfigures plot the dependance of the coverage $c$ on the user degree. For a given $x$, its corresponding $c$ is obtained by averaging all the users whose degrees are in the range of $[a(x^2-x),a(x^2+2)]$, where $a$ is chosen as $\frac{1}{2}\log 5$ [13]. The bottom two subfigures plot the relations between the average coverage $\bar{c}$ and the macro-step.
doi:10.1371/journal.pone.0079354.g003

$\langle RS \rangle$ of these methods are similar to MD in sparse networks. The parameters in these methods only slightly influence the results. A network manipulation method was proposed to solve the object cold-start problem by adding some virtual links to the network [15]. However, this method is also found less effective than the SLD method. This is because the virtual links inevitably contain some noise and the recommendation based on sparse data is very sensitive to the noise.

Additionally, we report the dependence of $\langle RS \rangle$ on the user degree and object degree in Fig. 5. The left two figures of Fig. 5 give the relationship between the user degree and $\langle RS \rangle$. One can see that $\langle RS \rangle$ of small-degree users who have collected few objects are improved greatly since these users' *coverage* of objects are increased significantly by the SLD. The right two figures of Fig. 5 show the relationship between the object degree and $\langle RS \rangle$. It can be seen that the SLD can improve $\langle RS \rangle$ of both the small-degree and large-degree objects.

Another interesting question is whether the accuracy of top-$L$ recommendation list will be improved the same as the ranking score by the SLD. The relation between the Recall and the number of macro-steps is presented in Fig. 6. For both datasets, we get the best performance when the macro-step is 2. However, when the macro-step exceeds 2, the Recall of both datasets starts to decrease. To uncover the reason, we study in detail the relationship between the top-$L$ accuracy and user degree and object degree, respectively. Since Recall is defined based on users,

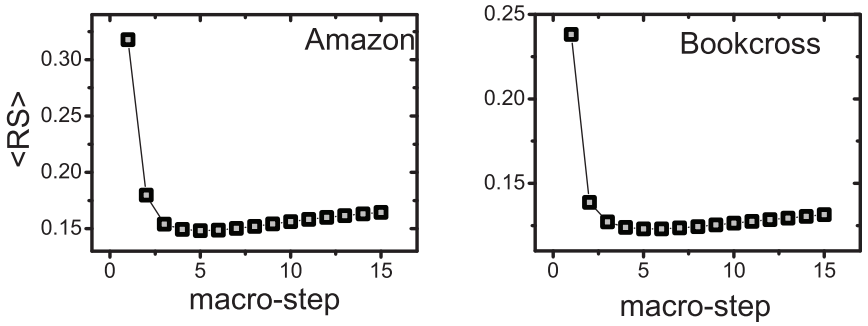


**Figure 4. The ranking score $\langle RS \rangle$ of the semi-local diffusion method in Amazon and Bookcross.** For both datasets, we obtain the lowest ranking score when the macro diffusion step is 5. Each data point is obtained by averaging over ten runs, each of which has an independently random division of training set and probe set.
doi:10.1371/journal.pone.0079354.g004

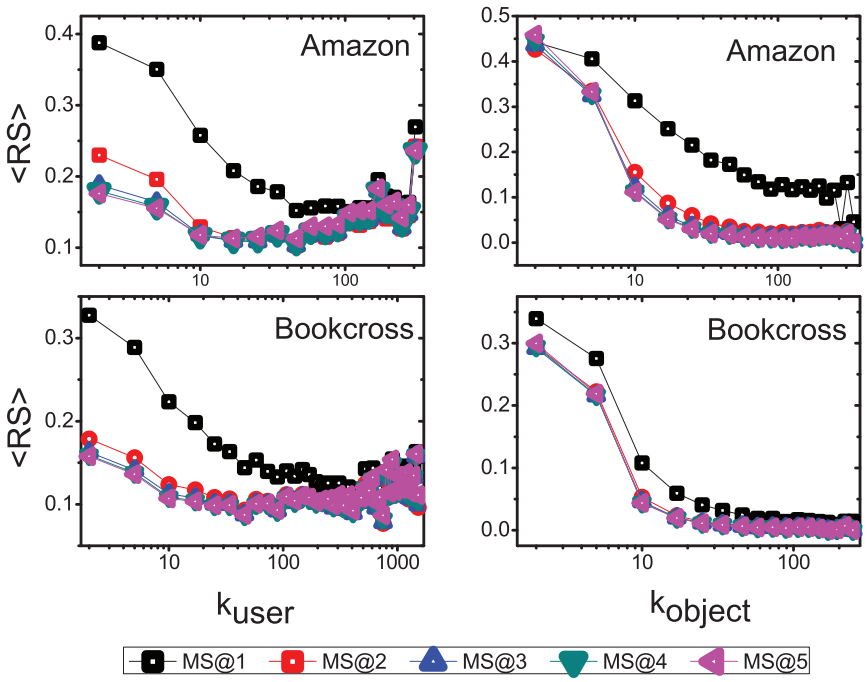


**Figure 5. Dependence of the ranking score $\langle RS\rangle$ on user degree and object degree.** The $MS@T$ means that $T$ is the macro-step of the diffusion. For a given $x$, its corresponding $\langle RS\rangle$ is obtained by averaging all the users (or objects) whose degrees are in the range of $[a(x^2-x),a(x^2+2)]$, where $a$ is chosen as $\frac{1}{2}\log 5$ [13]. Each data point is obtained by one run since the degree of a user and an item may change for different dataset divisions.
doi:10.1371/journal.pone.0079354.g005

it can be naturally used to measure the recommendation accuracy of the users with the same degree. When applied to objects, we define the object Recall as: $Re_\alpha(L)=d_\alpha(L)/E_\alpha$ where $E_\alpha$ is the number of users who selected object $\alpha$ in the probe set, and $d_\alpha(L)$ is the number of times that $\alpha$ appears in these $E_\alpha$ users' recommendation lists. The Recall of the objects with the same degree is obtained by simply averaging $Re_\alpha(L)$ of these objects. The top two subfigures of Fig. 7 show *Recall* of the users whose degrees are no larger than 5 . It can be seen that the accuracy of top-20 recommendation lists of those inactive users are improved considerably by the SLD if the macro-step of diffusion is less than 5. The best macro-step is 3 for Amazon and 4 for Bookcross, respectively. If the macro-step of diffusion exceeds 5, the *Recall* of those users starts to decrease. The bottom two subfigures of Fig. 7 give the *Recall* of the users whose degrees are no smaller than 20. It shows that the *Recall* decreases monotonously with macro-step. In addition, we plot the relationship between *Recall* and the object degree in Fig. 8. It shows that the SLD method tends to improve the *Recall* of large-degree objects. Generally speaking, small degree users incline to select popular items [28]. However, since the small degree users only have limit number of links, the original 3-step diffusion cannot reach the relevant popular items for them. On the other hand, the SLD method effectively increases the diffusion coverage and discover the most relevant popular items for these small degree users. This is of great importance from practical point of view since these new/inactive users are very sensitive to the quality of recommendation and poor quality may lead to losing them from the website.

Our result above shows that the high order diffusion resources may play different role in the recommendation for users and objects with different degrees. Therefore, the information of the high order diffusion should be used in a personalized way. Accordingly, we propose two extended recommendation methods: the user-based semi-local diffusion method (U-SLD for short) and the object-based semi-local diffusion method (O-SLD for short). We denote $\vec{f}^{(1)},\vec{f}^{(2)},\ldots,\vec{f}^{(n)}$ as the final resource vectors after *1*, *2*, …, $n$ macro-steps of diffusion, respectively. $\vec{f}^{(n)}$ can be easily calculated by $\vec{f}^{(n)}=W\vec{f}^{(n-1)}=W^n\vec{f}$. Given the target user $u$, the user-based semi-local diffusion method is to combine these $n$ resource vectors based on $u$'s degree. Mathematically, the final score of object $\alpha$ reads

$$F_\alpha^u=f_\alpha^{(1)}+\sum_{i=2}^{n}\frac{1}{(K-k_u)^\theta}f_\alpha^{(i)}, \tag{5}$$

where $k_u$ is $u$'s degree, $K=max(k_u)+1$ and $\theta$ is a free parameter to tune the weight of $\vec{f}^{(i)}$ $(i\geq 2)$ based on $u$'s degree. If $\theta>0$, the second term will play a more significant role when recommending objects for large-degree users, and vice versa.

In the sparse dataset, the coverage of 3-step diffusion is very low. Even some popular items cannot be effectively reached by users. The object-based semi-local diffusion method accumulates those resources based on the object degree. The final score of object $\alpha$ computed by this method is

$$F_\alpha^u=f_\alpha^{(1)}+\sum_{i=2}^{n}\frac{1}{k_\alpha^\theta}f_\alpha^{(i)}. \tag{6}$$

If $\theta>0$, the second term will play a more significant role in calculating the score for small-degree items, and vice versa. We

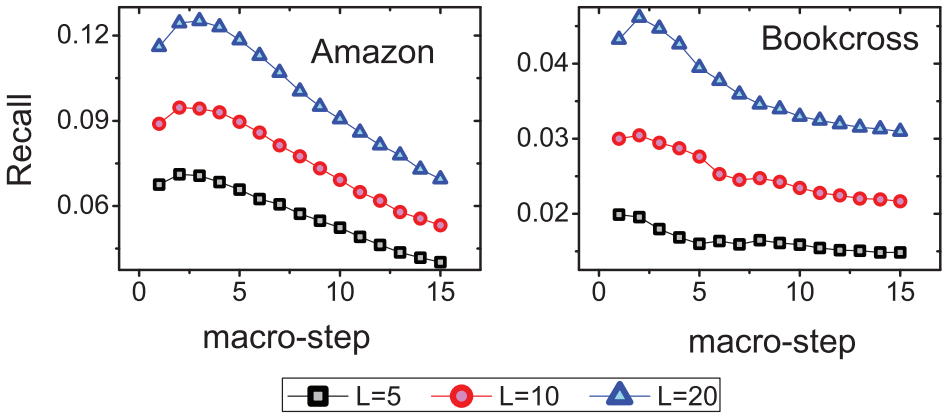


**Figure 6. The Recall of the semi-local diffusion method in Amazon and Bookcross.** For both datasets, we obtain the best performance when the macro-step of the diffusion is 2. Each data point is obtained by averaging over ten runs, each of which has an independently random division of training set and probe set.
doi:10.1371/journal.pone.0079354.g006

sort the vector $F^u$ in descending order and those objects with highest scores will be recommended to $u$. The results on Amazon and Bookcross are reported in Fig. 9 and the optimal parameters $\theta$ of algorithms discussed above are presented in Table 2. In order to balance the improvement on *ranking score* and *Recall*, we set $n=3$ in both U-SLD and O-SLD.

Actually, similar idea has been applied to eliminate the redundant correlations in dense datasets [29]. The method in [29] is called RENBI method and defined as

$$\vec{f'}=(W+\theta W^2)\vec{f}, \tag{7}$$

where the elements of matrix $W$ are defined by Eq. 1 with $\lambda=1$, $\vec{f'}$ and $\vec{f}$ is the final resource vector and the initial resource vector, respectively, and $\theta$ is a free parameter. In [29], the authors

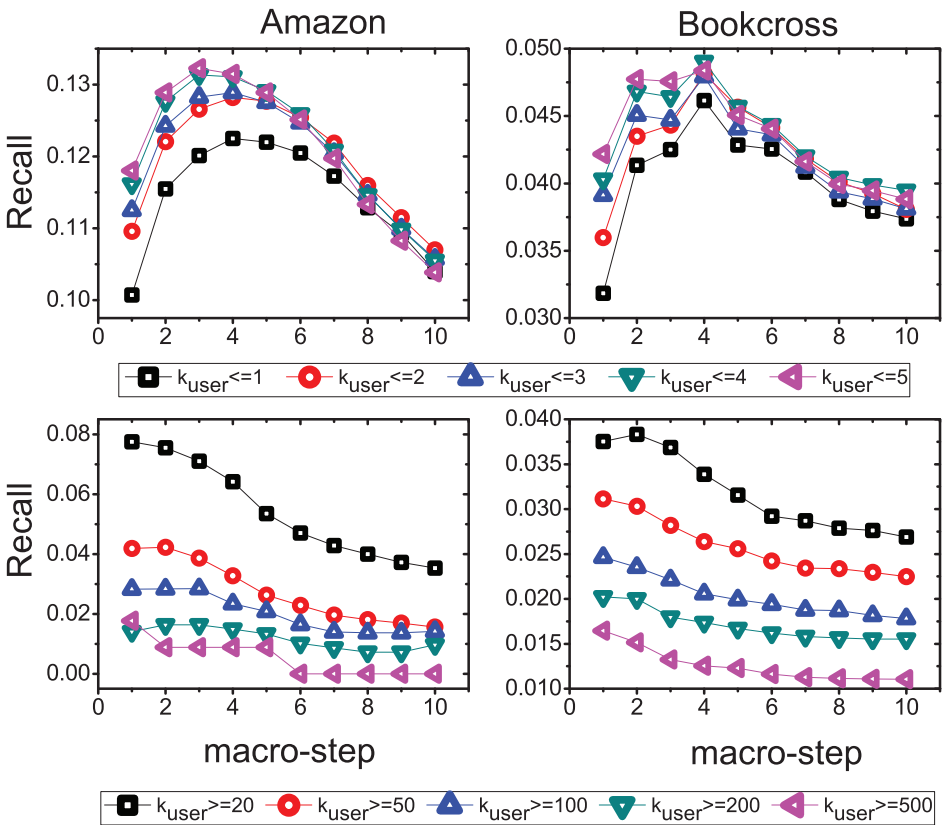


**Figure 7. Dependence of *Recall* on the diffusion macro-step.** The recommendation list length $L$ is set to 20. $k_{user}<=D$ means that we only consider the users whose degree is no larger than $D$. Each data point is obtained by one run since the degree of a user and an item may change for different dataset divisions.
doi:10.1371/journal.pone.0079354.g007

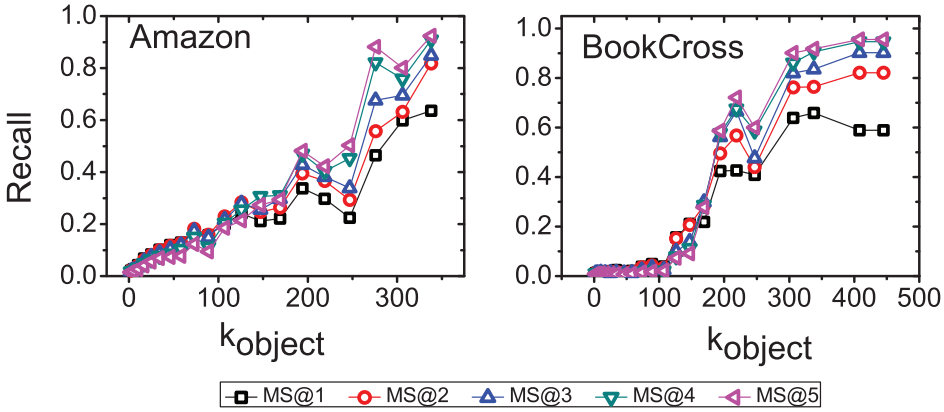


**Figure 8. The relationship between object degree and *Recall*.** The recommendation list length $L$ is set to 20. The $MS@T$ means that $T$ is the macro-step of the diffusion. Each data point is obtained by one run since the degree of a user and an item may change for different dataset divisions.
doi:10.1371/journal.pone.0079354.g008

focus on improving the accuracy and diversity of recommendation by eliminating the redundant information and they find that the optimal $\theta$ defined in Eq. 7 is negative. However, the information of high order diffusion is not redundant any more in sparse dataset. Moreover, the RENBI method is not personalized since the weight of high order diffusion resources is the same for all users. We will compare the U-SLD and O-SLD methods to the RENBI method.

The top subfigures of Fig. 9 show the results of *Recall* in Amazon and Bookcross. Clearly, the Recall of SLD is much higher than that of MD in both datasets. This is because the recommendation accuracy of small-degree users is significantly improved by SLD. Moreover, the RENBI method is also better than the MD method, but it is worse than the SLD. From the Table 2, we can also see that the optimal $\theta$ in Eq. 7 are 0.9 for Amazon and 0.7 for Bookcross, respectively. This is different from the result in ref [29] where the method is tested in dense data and the optimal $\theta$ is found to be negative. Our results indicate that the information of high order diffusion is in fact not redundant information in the sparse data. Both the U-SLD and O-SLD methods are better than the RENBI method in Recall. The improvement is due to the personalized use of the high order diffusion information. Finally, we can see that the O-SLD achieves the best *Recall* among these methods and the optimal $\theta$ defined in Eq. 6 is negative in both datasets from Table 2. That is to say, the information of high order diffusion should be considered more on the large degree items than small degree items. This is because small degree users inclines to select the popular items while these items cannot be effectively reached by one macro-step diffusion. Note that once those small degree users have selected many objects, we could then recommend diverse objects to them.

The bottom subfigures of Fig. 9 show the results of *ranking score* in Amazon and Bookcross. One can see that the ranking score of SLD method is much lower than that of MD. From the Table 2, it is shown that the optimal diffusion step is 5 in both datasets. RENBI also achieves a considerable improvement in ranking score compared to MD, but its ranking score is higher than that of SLD. The optimal $\theta$ of RENBI is also positive in both datasets. This supports again that the high order diffusion information is actually useful in enhancing the recommendation accuracy in sparse data. Although the ranking score of U-SLD and O-SLD method are slightly higher than the SLD method, these two methods enjoy a much better ranking score than RENBI. Taking together the results of ranking score and Recall, O-SLD seems to be the best recommendation algorithm in sparse data based on these training sets. It provides not only a good ranking of users' unselected objects but also an accurate top-$L$ recommendation list.

**Table 2.** The optimal parameter defined in algorithms for *Recall* and *Ranking score*.

| **Amazon** | | | | | |
|---|---|---|---|---|---|
| | | **SLD-T** | **RENBI** | **U-SLD** | **O-SLD** |
| Recall | T | 2 | – | – | – |
| | $\theta$ | – | 2 | −0.9 | −0.3 |
| Ranking score | T | 5 | – | – | – |
| | $\theta$ | – | 2 | −1 | −0.5 |
| **Bookcross** | | | | | |
| | | **SLD-T** | **RENBI** | **U-SLD** | **O-SLD-** |
| Recall | T | 2 | – | – | – |
| | $\theta$ | – | 1.0 | −0.6 | −0.2 |
| Ranking score | T | 5 | – | – | – |
| | $\theta$ | – | 2 | −1 | −0.7 |

doi:10.1371/journal.pone.0079354.t002

## Discussion

The data sparsity problem is one of the biggest challenges in recommender systems. There are a large number of online users and objects with very few connections, which leads to the poor performance of many well-known recommendation algorithms. However, the data sparsity problem has not yet been systematically studied and not yet well addressed. Take the hybrid method [12] for example, one cannot get an improved recommendation accuracy when combining the mass diffusion and heat conduction algorithms. As a matter of fact, the data of most real online systems is much sparser than the data used in this paper. Therefore, solving the data sparsity problem is of great significance from the practical point of view.

In this paper, we propose a semi-local diffusion (SLD) method to solve the data sparsity problem in recommender systems. The results on two real online datasets indicate that our method significantly outperforms other well-known algorithms. Two personalized semi-local diffusion methods are also proposed which

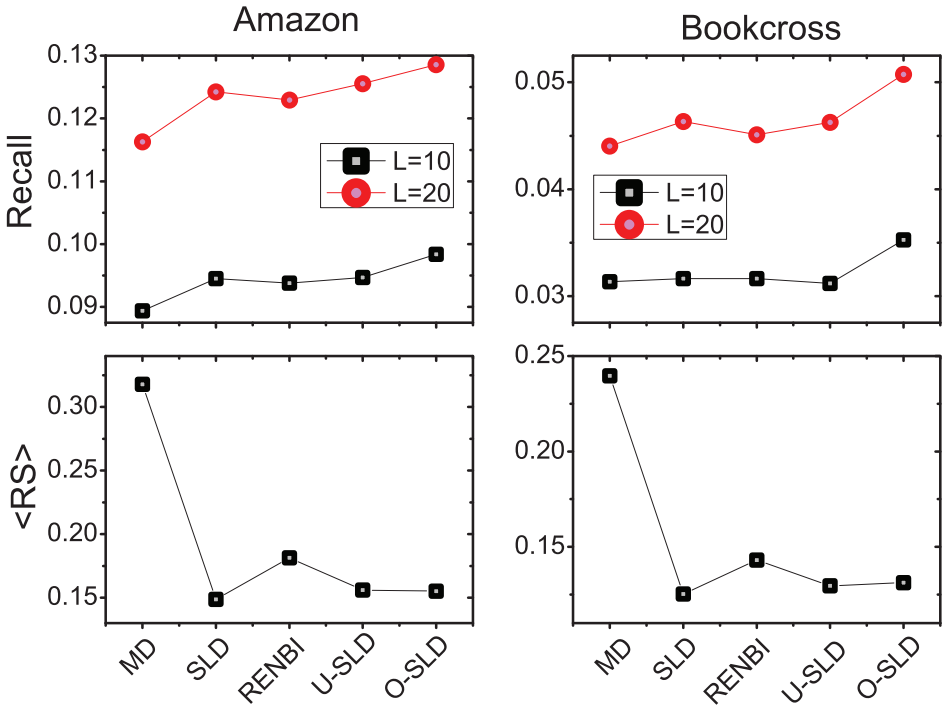


**Figure 9. The accuracy comparison of different algorithms.** Each data point is obtained by averaging over ten runs, each of which has an independently random division of training set and probe set.
doi:10.1371/journal.pone.0079354.g009

further improve the accuracy. Our analysis shows that the recommendation accuracy of small-degree users is greatly improved by the SLD method. In practical use, it can largely improve the experience of the new comers, so that more users will be attracted by the web site.

Finally, we remark that sparse online system are essentially different from the dense online system. Actually, most diffusion-based recommendation algorithms can be decomposed into two steps. The first step is to find all the relevant objects to the target user (i.e. objects covered by diffusion) and the second one is to rank these relevant objects. In the dense systems, the number of relevant objects is generally very large. Therefore, an effective recommendation algorithm in these systems should provide an accurate ranking of these relevant objects. However, the relevant objects in sparse systems are usually very limited and the objects the target user interested in might not be included in her/his relevant objects after 3-step diffusion. Accordingly, a more important issue for the recommendation algorithm in these systems should properly enlarge the number of relevant objects. Since the main task in designing recommendation algorithms in these two systems are different, all the algorithms and conclusions based on dense data should be rechecked in sparse data.

## Author Contributions

Conceived and designed the experiments: WZ AZ MSS YCZ. Performed the experiments: WZ AZ. Analyzed the data: WZ AZ. Contributed reagents/materials/analysis tools: WZ AZ. Wrote the paper: WZ AZ MSS YCZ.

## References


1. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17: 734–749.
2. Shang MS, Lü LY, Zeng W, Zhang YC, Zhou T (2009) Relevance is more significant than correlation: Information filtering on sparse data. EPL 88: 68008.
3. Zeng W, Shang MS, Zhang QM, Lü LY, Zhou T (2010) Can dissimilar users contribute to accuracy and diversity of personalized recommendation? Int J Mod Phys C 21: 1217–1227.
4. Pazzani MJ, Billsus D (2007) The adaptive web. Berlin, Heidelberg: Springer-Verlag, chapter Content-based recommendation systems. 325–341.
5. Maslov S, Zhang YC (2001) Extracting hidden information from knowledge networks. Phys Rev Lett 87: 248701.
6. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42: 30–37.
7. Hu YF, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. ICDM '08, pp.263–272.
8. Zhou T, Ren J, Medo M, Zhang YC (2007) Bipartite network projection and personal recommendation. Phys Rev E 76: 046115.
9. Zhang YC, Medo M, Ren J, Zhou T, Li T, et al. (2007) Recommendation model based on opinion diffusion. EPL 80: 68003.
10. Zhang YC, Blattner M, Yu YK (2007) Heat conduction process on community networks as a recommendation model. Phys Rev Lett 99: 154301.
11. Yang J, Kim J, Kim W, Kim YH (2012) Measuring user similarity using electric circuit analysis: Application to collaborative filtering. PLoS ONE 7: e49126.
12. Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, et al. (2010) Solving the apparent diversityaccuracy dilemma of recommender systems. Proc Natl Acad Sci USA 107: 4511–4515.
13. Lü LY, Liu WP (2011) Information filtering via preferential diffusion. Phys Rev E 83: 066119.
14. Liu JG, Zhou T, Guo Q (2011) Information filtering via biased heat conduction. Phys Rev E 84: 037101.
15. Zhang FG, Zeng A (2012) Improving information filtering via network manipulation. EPL 100: 58005.
16. Qiu T, Chen G, Zhang ZK, Zhou T (2011) An item-oriented recommendation algorithm on coldstart problem. EPL 95: 58003.
17. Zeng A, Yeung CH, Shang MS, Zhang YC (2012) The reinforcing influence of recommendations on global diversification. EPL 97: 18005.
18. Jamali M, Ester M (2009) Trustwalker: a random walk model for combining trust-based and itembased recommendation. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '09, pp.397–406.

19. Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems. RecSys '10, pp.135–142.
20. Ma H, Zhou DY, Liu C, Lyu MR, King I (2011) Recommender systems with social regularization. In: Proceedings of the fourth ACM international conference on Web search and data mining. WSDM '11, pp.287–296.
21. Yang SH, Long B, Smola A, Sadagopan N, Zheng ZH, et al. (2011) Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th international conference on World Wide Web. WWW '11, pp.537–546.
22. Yuan Q, Chen L, Zhao SW(2011) Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. In: Proceedings of the fifth ACM conference on Recommender systems. RecSys '11, pp.245–252.
23. Zhou T, Medo M, Cimini G, Zhang ZK, Zhang YC (2011) Emergence of scale-free leadership structure in social recommender systems. PLoS ONE 6: e20648.
24. Slanina F, Zhang YC (2005) Referee networks and their spectral properties. Acta Phys Pol B 36: 2797.
25. Ziegler CN, McNee SM, Konstan JA, Lausen G (2005) Improving recommendation lists through topic diversification. In: Proceedings of the 14th international conference on World Wide Web. WWW '05, pp.22–32.
26. Shi Y, Zhao X, Wang J, Larson M, Hanjalic A (2012) Adaptive diversification of recommendation results via latent factor portfolio. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. SIGIR '12, pp.175–184.
27. Zhang QM, Zeng A, Shang MS (2013) Extracting the information backbone in online system. PLoS ONE 8: e62624.
28. Shang MS, Lü LY, Zhang YC, Zhou T (2010) Empirical analysis of web-based user-object bipartite networks. EPL 90: 48006.
29. Zhou T, Su RQ, Liu RR, Jiang LL, Wang BH, et al. (2009) Accurate and diverse recommendations via eliminating redundant correlations. New J Phys 11: 123008.